Information and Communication Technologies

EPIWORK

Developing the Framework for an Epidemic Forecast Infrastructure

http://www.epiwork.eu

Project no. 231807

D 3.2 Prototype of the Epidemic Marketplace Platform with an initial set of epidemiological databases integrated available to project participants

---

Period covered:
Start date of project: February 1st, 2009
Due date of deliverable: January 31st 2010
Distribution: Consortium Only

Date of preparation: January 30th 2010
Duration:
Actual submission date: January 31st 2010
Status:

Project Coordinator: Alessandro Vespignani
Project Coordinator Organisation Name: ISI Foundation
Lead contractor for this deliverable: FFCUL

Work package participants

The following partners have taken active part in the work leading to the elaboration of this document, even if they might not have directly contributed writing parts of this document:

- Luís Filipe Lopes, FFCUL
- Fabrício Silva, FFCUL
- Francisco Couto, FFCUL
- Mário J. Silva, FFCUL

Change log

Version	Date	Amended by	Changes
1.0	2010-01-31	Mário J. Silva	Integrated input and made final revisions for delivery to the consortium for review.

D 3.2 Prototype of the Epidemic Marketplace Platform with an initial set of epidemiological databases integrated available to project participants

Mário J. Silva, Fabrício A. B. da Silva, Luis Filipe Lopes, Francisco M. Couto

LASIGE, University of Lisbon, Portugal
epiwork@di.fc.ul.pt

1. Introduction

This report describes the architecture and deployment status of the Epidemic Marketplace as it was available for Consortium use at the end of Month 12 of Epiwork, an e-Science platform for collecting, storing, managing and epidemic semantically annotated data collections for epidemic modellers.

In recent years, the availability of a huge flow of quantitative social, demographic and behavioural data spurred the interest on innovative technologies to improve disease surveillance systems, providing faster and better geo-referenced outbreak detection capabilities. These capabilities depend on the availability of fine-tuned models, which require accurate and comprehensive data. However, the increasing amount of data brings in the problem of data integration and management. New solutions are needed to ensure that data are correctly stored, managed and made available to the scientific community. For instance, digital repository systems were developed to provide the framework for creation, management, and preservation of existing and evolving forms of digital content [1]. These systems are only effective if they 1) collect, preserve and provide data in multiple formats; 2) provide user access management features; 3) organize data according to multiple dimensions, including subject, relevance and accuracy; 4) support metadata annotation to describe the data; 5) involve the community in an active way.

The Epidemic Marketplace prototype is available at http://epiwork.di.fc.ul.pt/.

2. Epidemic Marketplace Requirements

The architectural requirements of the Epidemic Marketplace are directly related to the objectives of the Epiwork project and have been defined according to the feedback from its partners. Although there are a number of projects that retrieve epidemic data and make them available to users, such as Healthmap [8], GPHIN [9], MedISys [11] and GIDEON [10], the set of requirements of the Epidemic Marketplace is unique and differentiates our platform from previous projects. The main requirements of the Epidemic Marketplace are listed below:

- **Support the sharing and management of epidemiological data sets**. Registered users should be able to upload annotated data sets, and a data set rating assessment mechanism should be available. The annotated data set will then compose a catalogue that will be available to users.
- **Support the seamless integration of multiple heterogeneous data sources.** Users should be able to have a unified view of related data sources. Data should be available from streaming, static and dynamic sources. All data retrieved by users or other services should be available through a common interface.
- **Support the creation of a virtual community for epidemic research.** The platform will serve as a forum for discussion that will guide the community into uncovering the necessities of sharing data between providers and modellers. Users will become active participants, generating information and providing data for sharing and collaborating online.
- **Distributed Architecture.** The Epidemic Marketplace should implement a geographically distributed architecture deployed in several sites. The distributed architecture should provide improved data access performance, improved availability and fault-tolerance.
- **Support secure access to data.** Access to data should be controlled. The marketplace should provide single sign on, distributed federated authorization and multiple access policies, customizable by users.
- **Support data analysis and simulation in grid environments.** The Epidemic Marketplace will provide data analysis and simulation services in a grid environment. Therefore, the Epidemic Marketplace should operate seamlessly with grid-specific services, such as grid security services, information services and resource allocation services.

- **Workflow.** The platform should provide workflow support for data processing and external service interaction. This requirement is particularly important for those services that retrieve data from the Epidemic Marketplace, process it, and store the processed data back in the marketplace, such as grid-enabled data analysis and simulation services.

3. Epidemic Marketplace Architecture and Deployment

The Epidemic Marketplace can be defined as a *distributed virtual repository,* a platform supporting *transparent*, seamless access to distributed, heterogeneous and redundant resources [2][3]. It is a *virtual* repository because data can be stored in systems that are external to the Epidemic Marketplace, and it provides *transparent* access because several heterogeneities are hidden from its users. The Epidemic Marketplace is composed of a set of interconnected data management nodes geographically distributed, sharing common canonical data models, authorization infrastructure and access interfaces. Data can be either stored in one or more repositories or retrieved from external data sources using authorization credentials provided by clients. Data can also be replicated among repositories to improve access time, availability and fault tolerance. However, data replication is not mandatory; in several cases data must be stored in a single site due to, for instance, security constraints. It is worth noting, though, that any individual repository that composes the Marketplace will enable virtualized access to these data, once a user provides adequate security credentials.

An Epidemic Marketplace node has the following modules:

- **Repository**: stores epidemic data sets and an epidemic ontology to characterise the semantic information of the data sets.
- **Mediator:** a collection of web services that will provide access to internal data and external sources, based on a catalogue describing existing epidemic databases through their metadata using state-of-the-art semantic-web/grid technologies.
- **Collector**: retrieves information of real-time disease incidences from publicly available data sources, such as social networks; after retrieval, the collector groups the incidences by subject and creates data sets to store in the repository.
- **Forum:** allows users to post comments on integrated data from other modules, fostering collaboration among modellers;

Several open-source tools and open standards are being used in the Epidemic Marketplace implementation and deployment process. We selected Fedora Commons [1] and Muradora [4] for the implementation of the main features of the repository. Access control in the repository implements the XACML [7], LDAP [6] and Shibolleth [5] standards.

The mediator is currently under development. It will implement several OAI [13] standards, like ORE and PMH. Clients will be able to search and query datasets and corresponding metadata through a RESTfull interface, after an initial authentication procedure.

A preliminary prototype of the Data Collector is now collecting data from Twitter on a daily basis. A new version of the Data Collector is being implemented, and this new version will have a graphical user interface and the capability of collecting data both actively and passively from multiples sources. The user will be able to dynamically configure new data collection processes thorough the graphical interface and a number of pre-defined services.

The forum is currently available. It is implemented using phpBB[12] and is integrated with other modules of the Epidemic Marketplace.

Lisbon is the first site where an Epidemic Marketplace node has been deployed. We envision near-future node deployments in the sites of our partners in Netherlands and Italy. The total number of Epidemic Marketplace nodes will depend on strategic decisions to be made by the Epiwork participants as the project evolves. It is worth noting that the epidemic Marketplace is able to handle data from any communicable disease, depending on the need of users.

Hardware architecture - For the deployment at Lisbon we have analyzed the requirements listed in the previous section and decided for the use of two DELL servers, each server having two quad-core processors (Athlon) with 16GB of main memory and two 1TB disks. The computing servers are complemented with two Iomega network storage units, each unit having four 1TB disks in a RAID 5 configuration. Computing servers and network storage units are interconnected through 1Gb Ethernet links. This system provides a good level of redundancy and makes crash recovery possible in a quick and easy way. Also at first this setting allows the use of one server for testing purposes while keeping the other with a stable running version.

External connectivity is provided initially by a shared multi-gigabit link between University of Lisbon and GÉANT, the European network for research and high education. The access to GÉANT will assure connectivity between the main site of the Epidemic Marketplace and other Epiwork partners.

4. Datasets

The repository already contains several resources added, mostly to demonstrate the repository functionality and the metadata schema. Among these resources are datasets, web resources such as sites containing relevant epidemiological information, references to Institutions working in the epidemiological and public health areas and even documents such as technical reports or scientific articles.

At the moment, the access to this information requires registration and logging into to the system, which at the moment is only available to the Epiwork partners. Someone who is not registered can enter the repository and browse public collections but can not access data.

The repository includes datasets from the Data Collector [14], which contains data collected from the Twitter. These datasets are composed from messages with references to diseases. It also contains other datasets such as datasets containing cumulative cases of H1N1 in Australia and a dataset of the US Air Transportation System.

Other documents stored in the repository include a document of H1N1 vaccine dose plans.

Other resources such as descriptions of Institutions or web sites, contain only metadata, describing those resources and their location.

5. Conclusion

The current version already has several of the main features of the outlined architecture such as, for instance, data management and data sharing support, secured access to data, user forum and an initial version of the Data Collector. Currently, the Epidemic Marketplace is being populated with epidemic data collections. We are relying on the community to guide its iterative development. We will also make the full source code available as Open Source and encourage the development of extensions. In the next months, the WP3 team will be focusing on uploading information about epidemic resources and their data into the repository, and improving its usability. The public release of the Epidemic Marketplace is scheduled for Month 20.

6. References

[1] Lagoze C, Payette S, Shin E, Wilper C. Fedora: an Architecture for Complex Objects and their Relationships. Inter. J on Digital Libraries Vol. 6, no. 2, pp 124-138, 2006.

[2] Kuliberda K, Blaszczyk P, Balcerzak G, Kaczmarski K, Adamus R, Subieta K. Virtual Repository Supporting Integration of Pluginable Resources. Proc. IEEE 17th Intern. Conf. on Databases and Expert Systems Applications (DEXA'06), 2006.

[3] Ohno-Machado L, Boxwala A, Ehresman J, Smith D, Greenes R. A Virtual Repository Approach to Clinical and Utilization Studies: Application in Mammography as Alternative to a National Database. Proc. 1997 AMIA Annual Symposium, 1997.

[4] Nguyen C, Dalziel J. Muradora: A Turnkey Fedora GUI Supporting Heterogeneous Metadata, Federated Identity, and Flexible Access Control. In: Proc. Third Intern. Conf. on Open Repositories, 2008.

[5] Shibboleth web site. http://shibboleth.internet2.edu/. Accessed Aug. 21, 2009.

[6] Tuttle S, Ehlenberger A, Gorthi R, Leiserson J, Owen N, Ranahandola S, Storrs M, Yang C. Understanding LDAP design and implementation. IBM International Technical Support Organization, 2nd ed., 2004.

[7] OASIS - eXtensible Access Control markup Language (XACML) web site. http://www.oasis-open.org/committees/xacml/charter.php/. Accessed Aug. 21, 2009.

[8] Brownstein, JS, Freifeld, CC.HealthMap: the development of automated real-time internet surveillance for epidemic intelligence. Euro Surveill 12: E071129 071125. Retrieved February 28, 2008 from http://www.eurosurveillance.org/ew/2007/071129.asp#5.

[9] Mawudeku A, Blench M. Global Public Health Intelligence Network (GPHIN). Proc. 7th Conf. of the Association for Machine Translation in the Americas. Retrieved Feb. 11, 2008 from www.mt-archive.info/MTS-2005-Mawudeku.pdf.

[10] GIDEON web site. http://www.gideononline.com/. Accessed Aug. 21, 2009.

[11] MedISys web site. http://medusa.jrc.it/. Accessed Aug. 21, 2009.

[12] phpBB web site. http://www.phpbb.com/. Acessed Jan. 31, 2010.

[13] Open Archives Initiative web site. http://www.openarchives.org/. Acessed Jan. 31, 2010...

[14] Lopes, LF, JM Zamite, BC Tavares, FM Couto, F Silva, and MJ Silva. "Automated Social Network Epidemic Data Collector." INForum informatics symposium. Lisboa, 2009.